Homework 2

Managing Big Data

1. Briefly discuss the role of following Hadoop ecosystem technologies

1.1. Yarn

Yarn is “Yet Another Resource Negotiator”, which “manages the processing resources of the Hadoop cluster”. Its role in the ecosystem is to allocate and manage the resources, schedule the jobs and run processing frameworks in data processing.

1.2. Zookeeper

Zookeeper “is a highly available, high-performance coordination service.” Its role is to coordinate many other roles in Hadoop ecosystem “by building blocks that can be used to build a large class of coordination data structures and protocols.”

1.3. Oozie

Oozie is a “workflow engine” that “runs jobs at specific times when data is present in a directory”. Its role is to schedule and run workflows of jobs, which can be Hive queries, Pig scripts, etc.

1.4. Sqoop

Sqoop is a tool to transfers data between HDFS and RDBMS. Its role is “a library and a runtime environment for efficiently moving data between relational databases and HDFS”.

1.5. Hue

Hue “provides a Web front-end to a Hadoop” and “access control for the cluster by requiring users to log in before they can use the system”. Its role is a “web-based user interface for Hadoop”.

Reference: ISOM 671 - 13 - Hadoop I.pdf, ISOM 671 - 14 - Hadoop II (MapReduce).pdf, ISOM 671 - 15 - Hadoop III.pdf.

2. Submit the list of pig script commands in text file (e.g. nyse.pig) file.

File: nyse.pig

Code:

dividends = LOAD 's3://hm3xiao/work1/NYSE\_dividends' as (org:chararray, company:chararray, dmy:chararray, dividend:float);

daily = LOAD 's3://hm3xiao/work1/NYSE\_daily' as (org:chararray, company:chararray, dmy:chararray, open:float, high:float, low:float, close:float, code:int, adjClose:float);

/\* 2.2.2. Join daily and dividends on stock and date \*/

df222 = JOIN dividends BY (company, dmy), daily BY (company, dmy);

/\* 2.2.3. Calculate dividend/close\_price \*/

df223 = foreach df222 generate \*, dividend/close as divOverClose;

/\* 2.2.4. Find stock ticker and date for minimum and maximum value of dividend/close\_price \*/

df224Min = order df223 by divOverClose asc;

df224Min = limit df224Min 1;

df224Min = FOREACH df224Min GENERATE dividends::company, dividends::dmy;

dump df224Min;

df224Max = order df223 by divOverClose desc;

df224Max = limit df224Max 1;

df224Max = FOREACH df224Max GENERATE dividends::company, dividends::dmy;

dump df224Max;

3. Submit the above hive queries in text file (e.g. nytaxi.hql) file.

File: nytaxi.hql

Code:

-- 3 Submit the above hive queries in text file (e.g. nytaxi.hql) file

-- 3.1. Upload tripdata.csv from canvas to a folder (hive) in your S3 bucket

-- 3.2. Create external table nyTaxi with

hive

create table nyTaxi(

VendorID int,

lpep\_pickup\_datetime string,

lpep\_dropoff\_datetime string,

store\_and\_fwd\_flag string,

RatecodeID int,

PULocationID int,

DOLocationID int,

passenger\_count int,

trip\_distance double,

fare\_amount double,

extra double,

mta\_tax double,

tip\_amount double,

tolls\_amount double,

ehail\_fee double,

improvement\_surcharge double,

total\_amount double,

payment\_type int,

trip\_type int

)

row format delimited FIELDS TERMINATED BY ','

lines terminated by '\n'

LOCATION 's3://hm3xiao/hive/'

tblproperties("skip.header.line.count"="1");

LOAD DATA INPATH 's3://hm3xiao/tripdata/tripdata.csv' overwrite INTO TABLE nyTaxi;

select \* from nyTaxi;

-- 3.3. Get distinct rate\_code\_id from the table

select distinct RatecodeID from nyTaxi;

-- 3.4. Show all rows/columns where rate\_code\_id = 1

select \* from nyTaxi where RatecodeID = 1;